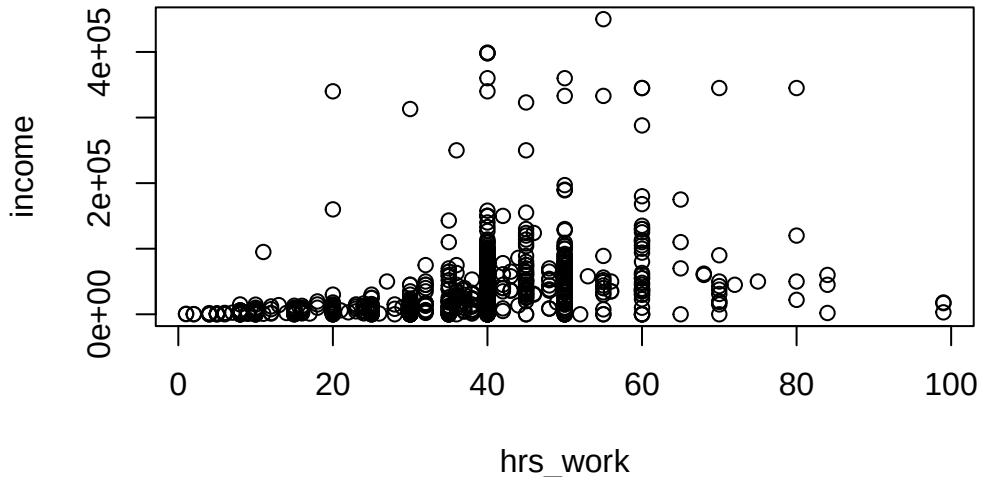
Exam 3

Mark Schulist

```
source("https://www.openintro.org/data/R/acs12.R")
```

a.

```
plot(income ~ hrs_work, data = acs12)
```



b.

```
model <- lm(income ~ hrs_work, data = acs12)
summary(model)
```

```
Call:
lm(formula = income ~ hrs_work, data = acs12)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-121855  -23313   -9189    7245  386372

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12905.5     4983.1   -2.59  0.00975 **
hrs_work      1391.5      123.6   11.25  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51650 on 957 degrees of freedom
  (1041 observations deleted due to missingness)
Multiple R-squared:  0.1169,    Adjusted R-squared:  0.116
F-statistic: 126.7 on 1 and 957 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 1391.5x - 12905.5$$

c. The intercept is -12905.5, which represents the number of dollars someone can expect to make if they work zero hours per week. This value does not make much sense as someone cannot earn negative money, and working zero hours should earn zero dollars.

d. The slope is 1391.5, which represents the additional dollars of income someone can expect to get for every additional hour worked each week.

e.

```
r_squared <- cor(acs12$income, acs12$hrs_work, use = "complete.obs") ^ 2
r_squared
```

```
[1] 0.1168822
```

11.7% of the variation in someone's income can be explained by the number of hours they work each week.

f.

H_0 : There is not a linear relationship between the number of hours worked and someone's income

H_a : There is a linear relationship between the number of hours worked and someone's income

The slope has a p-value of nearly zero (as shown in the summary), so we reject the null hypothesis and can conclude that there is a linear relationship between the number of hours worked each week and someone's income.

f.

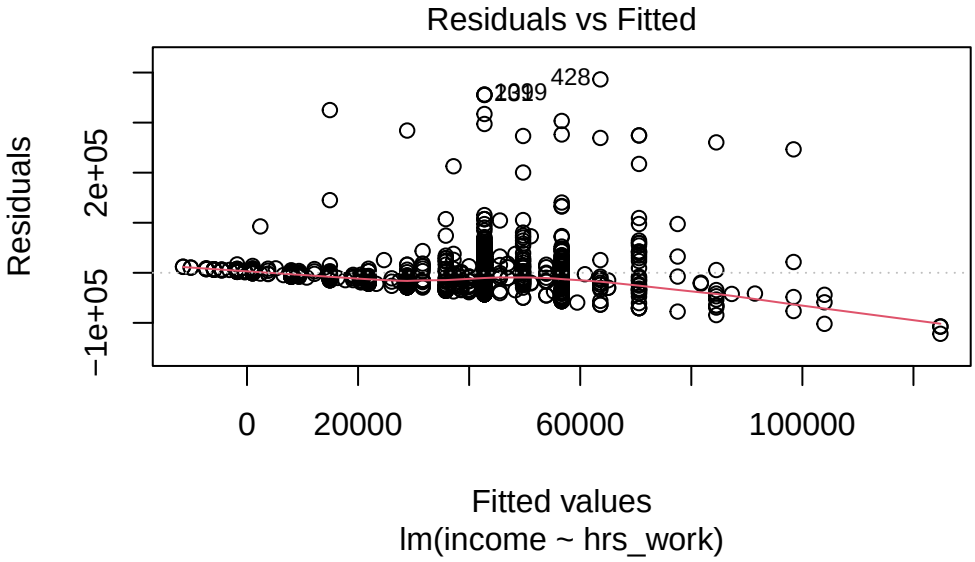
```
new_hrs <- data.frame(hrs_work = 40)
predict(model, newdata = new_hrs, interval = "confidence")
```

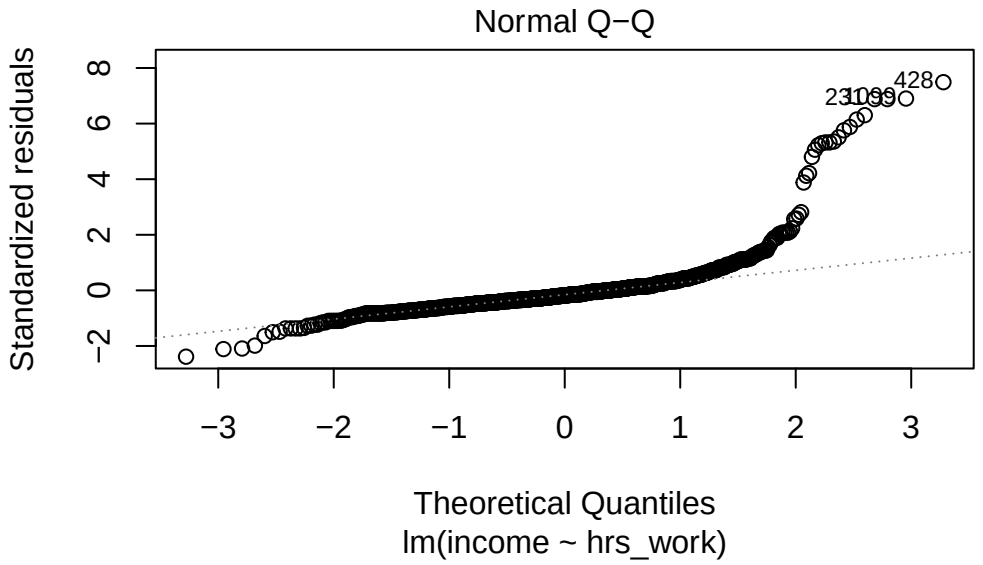
```
       fit      lwr      upr
1 42755.28 39445.29 46065.28
```

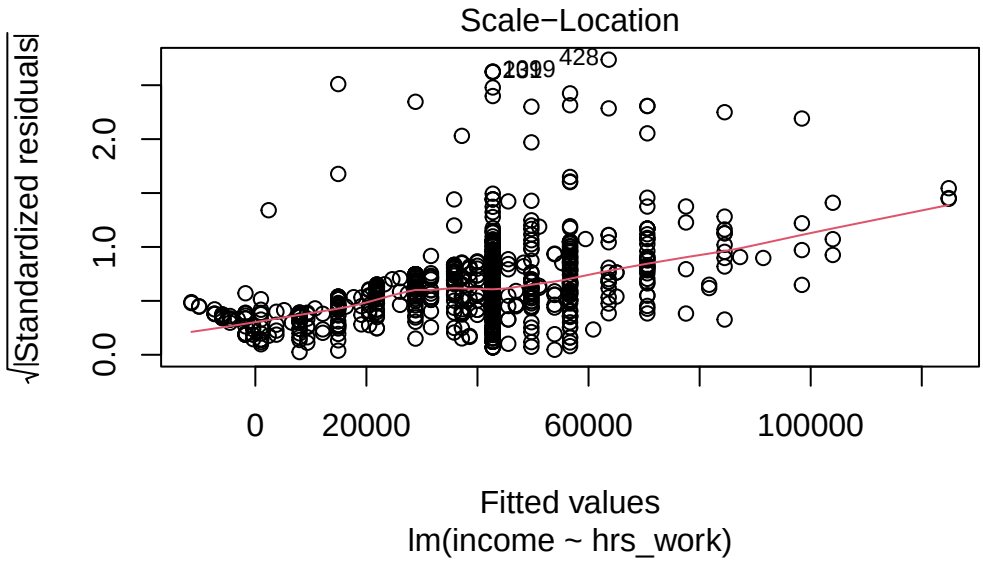
We estimate that the average income of someone working 40 hours a week is $42755.28. The 95% confidence interval is (39445.29, 46065.28).

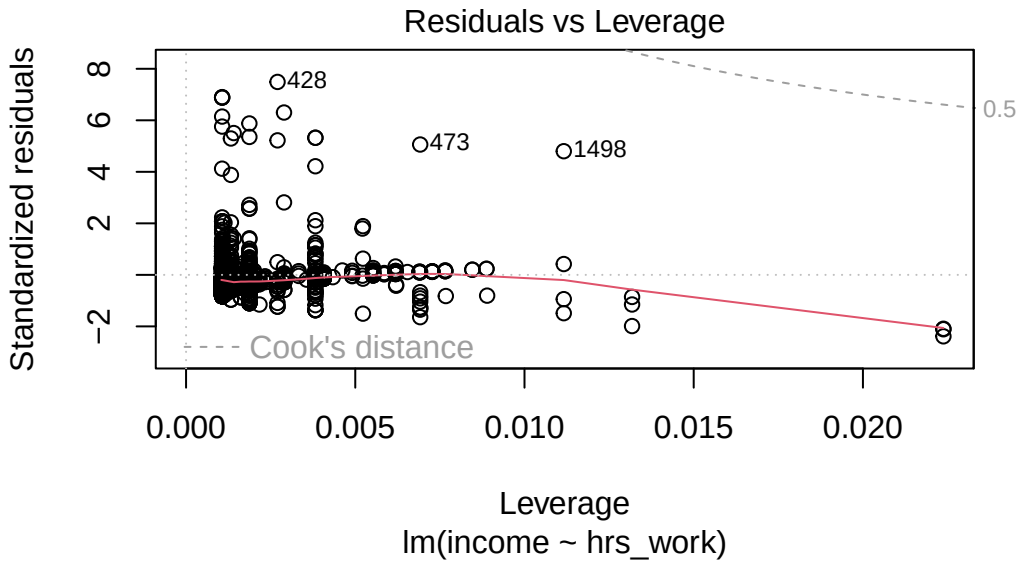
h.

```
plot(model)
```




Normal Q−Q
Standardized residuals
−2
0
2
4
6
8
428
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(income ~ hrs_work)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
428
0
20000
60000
100000
Fitted values
lm(income ~ hrs_work)



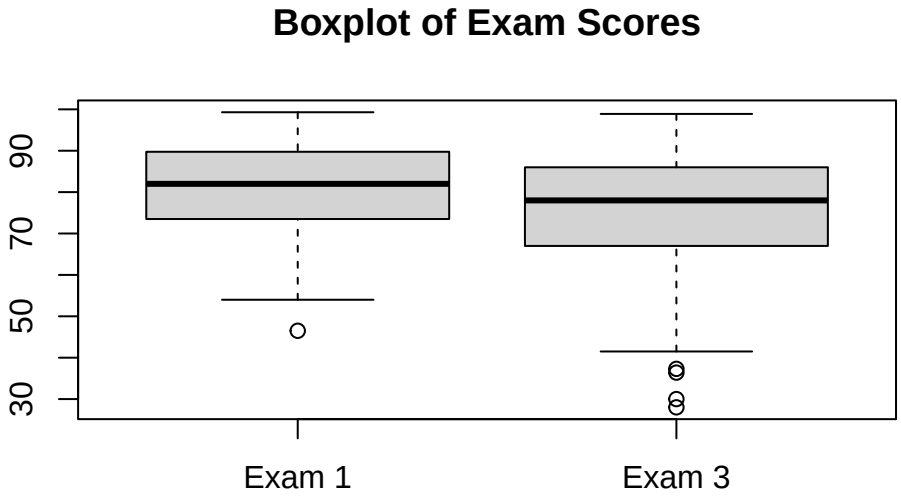
This test is not very valid, as shown by both the residuals vs. fitted plot and the QQ plot. The residuals are not evenly spread around the best fit line, with a lot of the points with positive residuals being further away than the negative residuals. The edges of the QQ plot are not close to the ideal line, which shows that the residuals are not normally distributed, as required by this model.

2.

```
source("https://www.openintro.org/data/R/exam_grades.R")
```

a.

```
boxplot(exam_grades$exam1, exam_grades$exam3,
        names = c("Exam 1", "Exam 3"),
        main = "Boxplot of Exam Scores"
        )
```



b.

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

```
t.test(exam_grades$exam1, exam_grades$exam3, paired = T)
```

```
    Paired t-test

data:  exam_grades$exam1 and exam_grades$exam3
t = 5.176, df = 231, p-value = 4.922e-07
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 3.281845 7.315948
sample estimates:
mean difference
       5.298897
```

The p-value is far below 0.05, so we reject the null hypothesis. We have evidence to show that the mean test scores between Exam 1 and 3 are different.

c. Looking at the boxplots, the data appear to be normally distributed. There is a small amount of left skew, but the same size is large enough that the test will be robust to this skew.

3.

```
source("https://www.openintro.org/data/R/ppp_201503.R")
```

a.

$$H_0 : \text{There is no association with political party and opinion on tax raises}$$

$$H_a : \text{Political party and opinion on tax raises are associated}$$

```
obs <- table(ppp_201503)
t <- chisq.test(obs)
t
```

```
    Pearson's Chi-squared test

data:  obs
X-squared = 139.41, df = 4, p-value < 2.2e-16
```

Our p-value is essentially 0, so we reject the null hypothesis. We have evidence to conclude that there is an association between political party and opinion on tax raises.

b.

```
t$expected
```

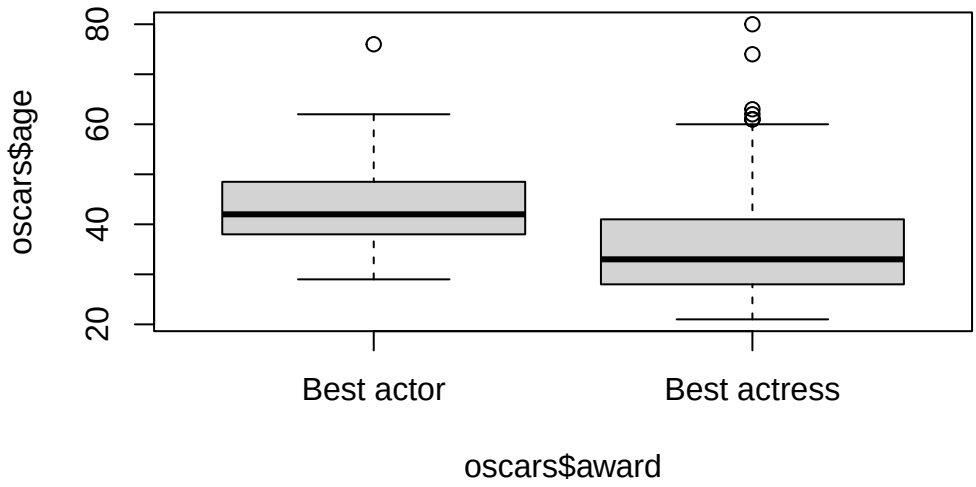
```
              taxes
party          Not sure Raise taxes on the poor Raise taxes on the rich
  Democrat      74.69175                21.96816                179.3401
  Indep / Other 48.71201                14.32706                116.9609
  Republican    63.59624                18.70478                152.6990
```

Our expected values are all above 5, so the chi-squared test is valid.

4.

```
source("https://www.openintro.org/data/R/oscars.R")
```

```
boxplot(oscars$age ~ oscars$award)
```



It is pretty clear that actors are older when receiving an Oscar than actresses. The median, as well as every quantile, are larger for actors than actresses.

b.

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

```
t.test(oscars$age ~ oscars$award)
```

```
    Welch Two Sample t-test

data:  oscars$age by oscars$award
t = 4.9178, df = 167.83, p-value = 2.072e-06
alternative hypothesis: true difference in means between group Best actor and group Best acti
95 percent confidence interval:
  4.547767 10.647885
sample estimates:
  mean in group Best actor mean in group Best actress
                  43.84783                   36.25000
```

The p-value is far below 0.05, so we reject the null hypothesis. We have evidence to suggest that true difference in means ages of actors and actresses when they get an Oscar is not equal to zero.

c. The two distributions (actors and actresses' ages) are mostly normal, with some right skew overall. Because of the large sample size (92 for each), the t test will be robust to this deviation from the assumptions.

d. If the conditions did not hold, we should use a nonparametric test such as a Wilcoxon Rank Sum Test. This test compares the medians (technically the entire distribution) and does not have the same conditions as parametric tests. This test only requires that the data are independent, not that they are normal.